GenBank, RefSeq, TPA and UniProt: What's in a Name?

The National Center for Biotechnology Information often is asked about the differences between its GenBank, RefSeq, and TPA databases and how they relate to the UniProt database. This document briefly describes the databases and elucidates some of the key differences.

GenBank

NCBI's GenBank database is a collection of publicly available annotated nucleotide sequences, including mRNA sequences with coding regions, segments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters.

GenBank is specifically intended to be an archive of primary sequence data. Thus, to be included, the sequencing must have been conducted by the submitter. NCBI does some quality control checks and will notify a submitter if something appears amiss, but it does not curate the data; the author has the final say on the sequence and annotation placed in the GenBank record. Authors are encouraged to update their records with new sequence or annotation data, but in practice records are seldom updated.

Records can be updated only by the author, or by a third party if the author has given them permission and notified NCBI. This delegation of authority has happened in a limited number of cases, generally where a genome sequence was determined by a lab or sequencing center and updating rights were subsequently given to a model organism database, which then took over ongoing maintenance of annotation.

Because GenBank is an archival database and includes all sequence data submitted, there are multiple entries for some loci. Just as the primary literature includes similar experiments conducted under slightly different conditions, GenBank may include many sequencing results for the same loci. These different sequencing submissions can reflect genetic variations between individuals or organisms, and analyzing these differences is one way of identifying single nucleotide polymorphisms.

GenBank exchanges data daily with its two partners in the International Nucleotide Sequence Database Collaboration (INSDC): the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). Nearly all sequence data are deposited into INSDC databases by the labs that generate the sequences, in part because journal publishers generally require deposition prior to publication so that an accession number can be included in the paper.

If part of a GenBank nucleotide sequence encodes a protein, a conceptual translation – called a coding region or coding sequence (CDS) – is annotated. A protein accession number (a "protein id") is assigned to the translation product and is noted on the GenBank record. This protein id is linked to a record for the protein sequence in NCBI's protein databases. In the UniProt database, described later, these sequences are contained in the TrEMBL (Translated EMBL) portion of the database.

Further information about GenBank is available in the NCBI Handbook; also see the GenBank overview at http://www.ncbi.nlm.nih.gov/Genbank/index.html.

RefSeq

The Reference Sequence (RefSeq) database is a curated collection of DNA, RNA, and protein sequences built by NCBI. Unlike GenBank, RefSeq provides only one example of each natural biological molecule for major organisms ranging from viruses to bacteria to eukaryotes. For each model organism, RefSeq aims to provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. RefSeq is limited to major organisms for which sufficient data is available (almost 4,000 distinct "named" organisms as of January 2007), while GenBank includes sequences for any organism submitted (approximately 250,000 different named organisms).

To produce RefSeq records, NCBI culls the best available information on each molecule and updates the records as more information emerges. A commonly used analogy is that if GenBank is akin to the primary research literature, RefSeq is akin to the review literature.

In some cases, creation of a RefSeq record involves no more than selecting a single good example from GenBank and making a copy in RefSeq, which credits the GenBank record. In other cases, NCBI in-house staff generates and annotates the records based on the existing primary data, sometimes by combining parts of several GenBank records. Also, some records are automatically imported from other curated databases, such as the SGDdatabase of yeast genome data and the FlyBasedatabase of Drosophila genomes (for a list of RefSeq collaborators see www.ncbi.nlm.nih.gov/RefSeq/collaborators). The approach selected for creating a RefSeq record depends on the specific organism and the quality of information available.

When NCBI first creates a RefSeq record, the record initially reflects only the information from the source GenBank record with added links. At this point, the record has not yet been reviewed by NCBI staff, and therefore it is identified as "provisional." After NCBI examines the record – often adding information from other GenBank records, such as the sequences for the 5'UTR and 3'UTR, and providing further literature references – it is marked as "reviewed."

RefSeq records appear in a similar format as the GenBank records from which they are derived. However, they can be distinguished from GenBank records by their accession prefix, which includes an underscore, and a notation in the "comment" field that indicates the RefSeq status. RefSeq records can be accessed through NCBI's Nucleotideand Proteindatabases, which are among the many databases linked through the Entrezsearch and retrieval system. When retrieving search results, users can choose to see all GenBank records or only RefSeq records by clicking on the appropriate tab at the top of the results page. Users also can choose to search only RefSeq records, or specific types of RefSeq

records (such as mRNAs), by using the “Limits” feature in Entrez. Further information about the database can be obtained at the RefSeq homepage.

Key Characteristics of GenBank *versus* RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common Records can contradict each other	Single records for each molecule of major organisms
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

TPA

The Third Party Annotation (TPA) database contains sequences that are derived or assembled from sequences already in the INSDC databases. Whereas DDBJ, EMBL and GenBank contain primary sequence data and corresponding annotations submitted by the laboratories that did the sequencing, the TPA database contains nucleotide sequences built from the existing primary data with new annotation that has been published in a peer-reviewed scientific journal. The database includes two types of records: experimental (supported by wet-lab evidence) and inferential (where the annotation is inferred and not the subject of direct experimentation).

TPA bridges the gap between GenBank and RefSeq, permitting authors publishing new experimental evidence to re-annotate sequences in a public database as they think best, even if they were not the primary sequencer or the curator of a model organism database. These records are part of the INSDC collaboration, and thus appear in all three databases (GenBank, DDBJ and EMBL).

Like GenBank and RefSeq records, TPA records can be retrieved through the Nucleotide section of Entrez. The TPA records can be distinguished from other records by the definition line, which begins with the letters "TPA," and by the Keywords field, which states "Third Party Annotation; TPA." Users can restrict their search to TPA data by selecting the database in the Properties search field or by adding the command “AND tpa[prop]” to their query. The database is significantly smaller than GenBank, with about one record for every 12,000 in GenBank. Details about how to submit data and examples of what can and cannot be submitted to TPA are provided on the TPA homepage.

UniProt

UniProt(Universal Protein Resource) is a protein sequence database that was formed through the merger of three separate protein databases: the Swiss Institute of Bioinformatics' and the European Bioinformatics Institute's Swiss-Prot and TrEMBL (Translated EMBL Nucleotide Sequence Data Library) databases, and Georgetown University's PIR-PSD (Protein Information Resource Protein Sequence Database).

Swiss-Prot and TrEMBL continue as two separate sections of the UniProt database. The Swiss-Prot component consists of manually annotated protein sequence records that have added information, such as binding sites for drugs. The TrEMBL portion consists of computationally analyzed sequence records that are awaiting full manual annotation; following curation, they are transferred to Swiss-Prot.

TrEMBL is derived from the CDS translations annotated on records in the INSDC databases, with some additional computational merging and adjustment. Given the very high rate of sequencing, and the effort it takes to do manual annotation, the Swiss-Prot component of UniProt is generally much smaller than the TrEMBL component. Because Swiss-Prot's manual annotation provides much additional information, NCBI's protein databases provide links to Swiss-Prot records, even if the sequence is the same as one or more INSDC translations.

Key Characteristics of UniProt *versus* GenBank and RefSeq

UniProt	GenBank and RefSeq
Produced by SIB, EBI & Georgetown U.	Produced by INSDC and NCBI
Protein data only	Protein and nucleotide data
Curated in Swiss-Prot, not in TrEMBL	Curated in RefSeq, not in GenBank